混合 CTC/attention 架构的端到端带口音普通话识别 *

杨　威，胡　燕†

(武汉理工大学 计算机科学与技术学院，武汉 430000)

摘　要：针对普通话语音识别任务中的多口音的识别问题，提出了链接时序主义(connectionist temporal classification，CTC)和多头注意力(MultiHead attention)的混合端到端模型，同时采用多目标训练和联合解码的方法。实验分析发现随着混合架构中链接时序主义权重的降低和编码器层数的加深，混合模型在带口音的数据集上表现出了更好的学习能力，同时训练一个深度达到 48 层的编码器-解码器架构的网络，生成的模型表现了超过之前所有端到端模型，在数据堂开源的 200h 带口音数据集上达到了 5.6%字错率和 26.2%句错率。实验证明了本文提出的端到端模型超过一般端到端模型的识别率，在解决带口音的普通话识别上有一定的先进性。

关键词：口音；混合 CTC/Attention 的端到端模型；多头注意力；链接时序主义；语音识别

中图分类号：TP912.34　　doi: 10.19734/j.issn.1001-3695.2020.02.0036

Hybrid CTC/attention architecture for end-to-end multi-accent mandarin speech recognition

Yang Wei, Hu Yan†

(*School of Computer Science & Technology*, *Wuhan University of Technology*, *WuHan Hube* 430000, *China*)

Abstract: To improve the performance of multi-accent Mandarin speech recognition task, this paper present a method for hybrid end-to-end automatic speech recognition(ASR) by combining Connectionist Temporal Classification (CTC) and MutiHead Attention by using a multiobjective training and joint decoding. Our analysis shows that hybrid model with lower CTC weight and deeper encoder layers performance better learning capacity. And we trained a very deep models with up to 48 layers for encode-decoder Architecture, which outperform all previous end-to-end ASR approaches on Aidatatang 200h multi-accent dataset, achieve 5.6% Character Error Rate(CER) and 26.2% Sentence Error Rate(SER) . The experiment proves that the recognition rate of the end-to-end model proposed in this paper exceeds the general end-to-end model, and it has certain advancedness in solving the Mandarin recognition with accents.

Key words: accent; hybrid CTC/attention end-to-end model; multi-head attention; connectionist temporal classification; speech recognition

0　引言

随着人工智能技术的飞速发展，语音识别技术已经成为了智能设备的标配，成为人机交互的重要手段之一。在语音识别技术中，口音一直是语音识别技术的难点[1]。对于普通话而言，由于方言众多，大部分人的普通话极易受到当地方言的影响。有研究表明[2]，超过 79.6%的普通话使用者有口音，他们之中有 44%有严重的口音。

为了解决语音识别中的口音的问题，最早有学者提出使用词典自适应的方法[1,3]，简单来说就是根据方言的发音习惯扩展发音词典，但是这种方法会增加词典的混淆，因此并不会取得明显的提升。所以解决口音问题的关键集中在了声学模型上，最简单直接的方法就是为每一种口音都建立一个声学模型[4]。这种方式对于每一种需要建立声学模型的方言来说都需要大量的方言数据进行训练才能得到较高的识别率，最后使用语音决策树选出最好的模型。因此，最好的方式是建立一个统一的自适应声学模型能够准确的识别不同的口音。

针对口音问题建立自适应的声学模型已经有很多学者进行了很多的工作。最早，在高斯-隐马尔可夫模型上(Gaussian mixture density hidden Markov model，GMM-HMM)模型上有研究者提出极大似然线性回归(maximum likelihood linear regression，MLLR)的方法[5]。后续也有研究者提出结合极大似然线性回归和最大后验概率估计(maximum a posteriori estimation，MAP)的方法在带上海口音的普通话语音识别中取得了一定的提升[6]。随着深度神经网络在自动语音识别(automatic speech recognition，ASR)中的发展，有学者提出增加一层特定的口音判别层，并用 KL 散度(Kullback-Leibler divergence，KLD)的方法训练该项层口音模型，在深度神经网络(deep neural networks，DNN)上针对非英语母语说话人的语音识别上取得了显著的提升[7]。后来又有学者结合上述模型和增加说话人特定特征(i-vector)[8]特征输入到特定的口音判别层的方式，在带口音普通话的识别上取得了很大的提升效果[9]。也有研究者提出一种融合特征下基于卷积神经网络的说话人语音分割模型，为解决普通话说话人和四川话说话人的语音识别场景的任务提供了一定的研究价值[10]。

在最近的研究中，在语音识别系统中采用长短时记忆循环神经网络(long-short term memory recurrent neural networks，LSTM-RNN)[11]的表现可以媲美最新的基于隐马尔可夫模型的深度神经网络(deep neural network hidden Markov models，DNN-HMM)系统的识别效果。研究者在基于长短时记忆循环神经网络的语音识别模型上进行了很多研究，从各个方面显著的改善了普通话语音识别系统的整体性能。由于基于长短时记忆循环神经网络网络的优秀表现，在带口音的语音识别方向，有研究者使用一层深度神经网络作为口音依赖特征层过滤特定口音的特征，用双向长短时记忆循环神经网络(bidirectional long short-term memory recurrent Neural networks，

收稿日期：2020-02-02；**修回日期：**2020-03-26　**基金项目：**湖北省自然科学基金项目(2019CFC919)

作者简介：杨威(1994-)，男，硕士，主要研究方向为模式识别；胡燕(1965-)，女(通信作者)，教授，硕导，主要研究方向为信息检索、数据挖掘(huyan@whut.edu.cn).

BLSTM-RNN)训练自适应的声学模型，在多口音的普通话识别上也取得了一定的效果[12]。上述方法在训练过程中需要对不同方言信息进行单独的分类和训练，对于部分方言数据缺乏的情况很难取得较好的识别率。同时上述模型仍然是遵循传统的模块化建模的方案，需要进行多阶段的独立训练(声学模型，发音字典，语言模型)，还需要一些丰富的专业知识来进行超参数的配置[13,14]。

相反，端到端的模型可以直接进行语音特征到文本的转录，不包含任何一个中间件，直接建立语音到文本之间的映射，可以潜在的优化最终任务的所有部分，同时在不同的语言体系下也可以利用相同的框架训练直接训练。端到端模型已经在很多语音识别任务中表现出来可以媲美最先进的传统模型的识别效果[13~17]。端到端模型可以分为三种主要的基本模型架构，基于链接时序主义的模型(CTC)[13,14,17,18]借鉴了马尔可夫假设有效的解决序列的动态对齐问题，基于注意力机制的模型(attention)[19,20,23,24]通过一个注意力机制解决了声学帧和标签之间的对齐问题，基于混合架构的模型(CTC/attention)通过对上述两种模型进行联合训练和联合解码，在解码和是被上都取得了优于上述单独一种方式训练的效果[21,22]。研究者针对端到端模型鲁棒性差，固定长的语音帧造成的时频信息和部分高频信息的损失问题，提出了ResNet-BLSTM 网络，有效的降低了模型的字错率[25]。

本文提出结合多头注意力(multi-head attention)的编码器-解码器模型[24,25]和链接时序主义(CTC)的混合端到端架构模型，采用联合训练和联合解码的方式[20,21]。在模型中，加深编码器-解码器网络的深度并训练了一个非常深的编码器-解码器网络，研究深层次网络对于语音识别的影响，同时在浅层的编码器-解码器网络中设置不同随机失活比率，研究带口音数据集的训练过程。

1　混合 CTC/attention 模型

本文提出的模型，如图 1，本文提出了一个基于多头注意力和链接时序主义的深层编码器-解码器网络，在混合架构中，链接时序主义(CTC)和多头注意力(multi-head attention)联合训练和联合打分，同时通过加深编码器和解码器网络层的方式进一步提升了识别率。

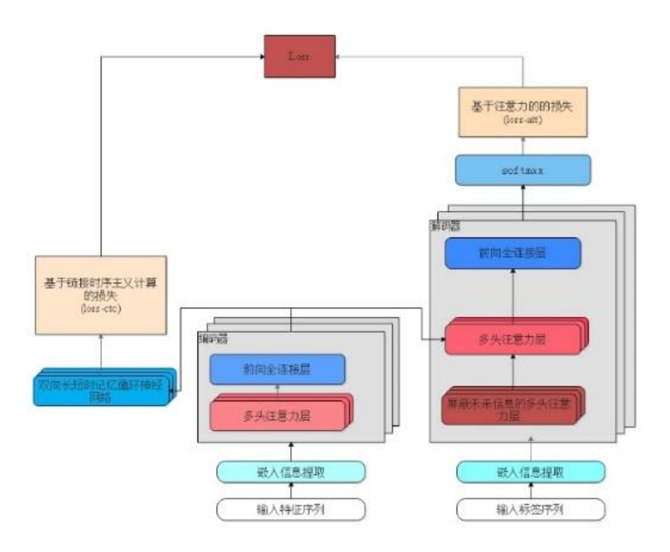


图 1　基于多头注意力和链接时序主义的深层编码器-解码器架构

Fig. 1　Deep encoder-decoder architecture based on multi-head attention and connectionist temporal classification

对于端到端的语音识别任务，目的是通过一个单一的网络，对于输入序列 $x=(x_1,\cdots,x_T)$，计算所有输出标签序列 $y=(y_1,\cdots,y_U)$ 的概率，通常 $U \le T$，$y_u \in L$，L 是有限字符集的集合，并输出最大概率的标签序列。即：

$$y^* = \arg\max_y P(y|x) \tag{1}$$

1.1　链接时序主义(CTC)

基于链接时序主义训练的网络如图 2 所示，在 CTC 训练中，定义中间标签序列 $\pi=(\pi_1,\cdots,\pi_T)$，在序列 π 中允许重复的标签存在，并插入一个空白标签 <-> 作为分隔符，即 $\pi_t \in L \cup \{<->\}$。假设 y' 是 y 增加了分隔符之后的扩展，例如 $y=(n,\ i,\ h,\ a,\ o)$，$y'=(-,n,-,i,-,h,-,a,-,o,-)$，定义一个多对一的映射 $B:L'->L^{\le T}$，其中 $L^{\le T}$ 是可能的中间标签序列 π 的输出集合，可以得到输出标签的 $P(y|x)$ 概率:

$$P(y|x) = \sum_{\pi \in B^{-1}(y')} P(\pi|x) \tag{2}$$

图 2　链接时序主义(CTC)

Fig. 2　Connectionist temporal classification

对于 π，在输入序列 x 的每个时间步 t 都要计算对应的输出 π_t，假设每个时间步之间的输出序列是有条件的独立的，可以得到式(3):

$$P(\pi|x) \approx \prod_{t=1}^{T} P(\pi_t|\pi_1,\pi_2,\cdots,\pi_{t-1},x), \forall \pi \in L' \tag{3}$$

由式(1)~(3)可以定义 CTC 的损失函数负对数概率值:

$$\begin{aligned} L_{ctc}(S) &= -\ln p(y|x) \\ &= -\ln \sum_{\pi \in B^{-1}(y')} p(\pi|x) \\ &= -\sum_{(x,y')\in S} \ln \sum_{\pi \in B^{-1}(y')} p(\pi|x) \\ &= -\sum_{(x,y')\in S} \ln \sum_{\pi \in B^{-1}(y')} \prod_{t=1}^{T} q_t(\pi_t), \forall \pi \in L' \end{aligned} \tag{4}$$

在式(4)中，S 代表输入序列 x 和输出标签之间的映射，$q_t(\pi_t)$ 代表标签 π_t 在时间步 t 的概率。在式(4)中，在计算过程中需要穷举出所有时间步的所有的标签序列，在计算上需要 N^T 次迭代，N 代表标签的总数，计算量太复杂。CTC 算法[17]借鉴 HMM 的前后项算法，优化计算的速度：

$$P_{ctc}(y|x) = \sum_{t=1}^{T}\sum_{u=1}^{|y|} \frac{\alpha_t(u)\beta_t(u)}{q_t(\pi_t)} \tag{5}$$

在式(5)中，$\alpha_t(u)$ 代表第 u 个标签在时间步 t 的前向概率，$\beta_t(u)$ 代表第 u 个标签在时间步 t 的后向概率。

1.2　基于多头注意力的编码器-解码器

本文使用的编码器-解码器模型[20,21]受到 google 的 transformer 模型[26,27]的启发，采用基于多头注意力的多层编码器-解码器解码器模型，同时在每层输入前都添加一层随机失活层。编码器中的每一层都是由一个多头注意力层和一个全连接的前向网络组成；解码器中的每一层由一个屏蔽当前时刻之前的注意力信息的多头注意力层和计算编码器输入的注意力层，全连接的前向网络三层网络组成。在输入端。如图 3，为单层的编码器—解码器网络结构。

编码器根据输入序列 $x=(x_1,\cdots,x_T)$ 计算一个输出的中间序列 $z=(z_1,\cdots,z_T)$，解码器将序列 z 作为输入，在每个时间步都会更新输出标签 $y=(y_1,\cdots,y_U)$ 中的一个标签。在更新输出标签过程中，模型也会根据前面的标签作为输入来更新下一个输出标签。如下:

$$z_t = Encoder(x) \tag{6}$$

$$y_t = Decoder(z, y_{t-1}) \tag{7}$$

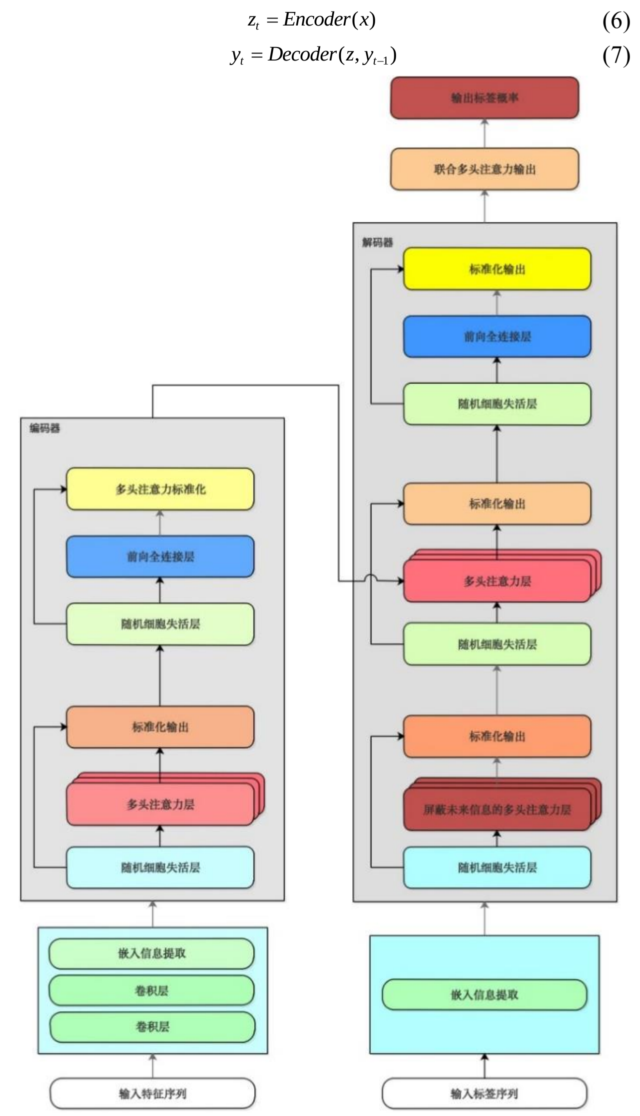


图 3　单层编码器-解码器网络结构

Fig. 3　Sequence-to-sequence keyword extension model based on attention mechanism

根据最大似然估计，可以最大化输出序列基于输入序列的条件概率作为损失函数:

$$\begin{aligned} L_{att} &= -\log p(y_1, y_2, \cdots, y_{T'} \mid x_1, x_2, \cdots, x_T) \\ &= \sum_{t'=1}^{T'} p(y_{t'} \mid y_1, y_2, \cdots, y_{t'-1}, z) \end{aligned} \tag{8}$$

从式(6)(7)可以看出，解码器的输入依赖于相同的序列 z，z 得到的是由序列 x 计算出的最终时间步的隐藏向量，而在实际的语言序列中，每个时刻标签的输出通常更加依赖于相近时刻的特征，所以在编码器-解码器模型中引入注意力机制。注意力机制在每个时刻都会输出一个结果 y_t，可以理解为将原来固定的中间序列 z 转换成会根据当前输出的动态变化的序列 z_t。

多头注意力机制首先对输入做线性转换初始化三个权重矩阵 Q、K、V:

$$Q = K = V = \text{Linear}(X) \tag{9}$$

然后根据点积的方式计算矩阵 Q，K 的相似度：

$$f(Q,K) = \frac{Q^T K}{\sqrt{d_k}} \tag{10}$$

对于解码器的输入，只能计算当前时刻和之前时刻输入特征的相似度，屏蔽未来信息，通常填充 0 将式(10)计算的矩阵变成下三角矩阵。然后用 softmax 函数做归一化输出，根据注意力分布计算加权平均:

$$Attention(Q,K,V) = soft\max(f(Q,K))V \tag{11}$$

为了避免过拟合问题，设置随机失活神经元，本文采用的是随机设置失活神经元的 dropout 函数，根据多头注意力的计算原则，将上面得到的结果重复计算 n 次，最终的结果做拼接，转换成线性输出：

$$Attention(Q,K,V) = soft\max(f(Q,K))V \tag{12}$$

$$MultiHead(Q,K,V) = Concat(head_1,\ldots,head_n)W^Q \tag{13}$$

进行完多头注意机制的计算，在编码器和解码器的每一层都要经过一个全连接的前向网络处理，包含两个线性变化和一个激活函数输出:

$$FFN(x) = (dropout(xW_1 + b_1))W_2 + b_2 \tag{14}$$

得到序列得分和注意力分布图。

1.3　多目标学习和联合解码

本文使用的混合模型受到[17,22]的启发，基于链接时序主义和基于多头注意力的模型联合训练，编码器由链接时序主义(CTC)和多头注意力模型(multi-head attention)共享，训练的过程中采用多目标学习(multiobjective learning，MTL)[22]的方式，相对于单一的数据驱动的注意力模型，利用了 CTC 的前-后向算法执行语音到标签之间的强制单音对齐，加速了对齐的过程。同时对比单一的 CTC 模型，由于注意力的目标是字符集的，CTC 的目标是序列级的，可以帮助提升 CTC 目标的准确性。使用交叉损失准则将和相结合的联合打分，提高了鲁棒性，由式(4)和(8)可以得出:

$$L_{MOL} = \lambda \log p_{ctc}(c \mid x) + (1-\lambda)\log p_{att}(c \mid x) \tag{15}$$

其中参数 $\lambda \in [0,1]$ 是一个插值权重。除了利用多目标学习训练网络，在混合模型中还有一个重要的步骤就是联合解码。

混合架构中，通常是使用集束搜索算法(beam search process)来执行解码过程，完整的集束搜索方法需要计算出每一种假设的得分。混合架构中需要联合链接时序主义的得分和注意力得分。在集束搜索中，利用了动态规划的思想，对于每一个时间步，都由前面可能的路径扩展，得到当前时间步的局部概率 g，这个局部概率由前面的路径决定，通过当前时间步可能的标签的概率 c，可以得到当前时间步的所有可能的假设 h，即当前标签 c 可不可能出现在当前时间步，得到假设后就可以通过式(15)计算联合得分，到下一个时间步继续扩展节点得到每个路径的打分，当预测到结尾字符时，将满足最小阈值要求的序列加入到最终序列中。比较所有可能序列的得分，取最高分为目标序列。本文在搜索解码过程中，采用多路搜索，每次选择概率最大的 k 个节点进行扩展，使每一步最多扩展 k 个节点，而不是随着时间步的增长呈现指数增加，可以很大程度的降低穷举搜索的复杂度。

$$\text{Score} = \arg\max_{h \in L'} \{\lambda \alpha_{ctc}(h,x) + (1-\lambda)\alpha_{att}(h,x)\} \tag{16}$$

2　实验

2.1　实验模型

本文采用的端到端模型为基于链接时序主义和基于多头注意力混合模型。对比采用的传统模块化模型为高斯-隐马尔可夫模型，时延神经网络模型(time-delay neural network, TDNN)和链式模型(chain)。对比采用的端到端模型为基于链接时序主义和基于位置信息的混合模型，基于多头注意力的模型。

2.2　数据

在本文的实验中，本文采用数据堂开源的 200h 带口音

普通话数据集，数据集中包含 30 万条口语化的句子，共 6408 人，其中男性 2999 人，女性 3301 人，录音人员分布于广东、福建、山东等 34 个省级行政区域。传统的模块化模型在 Kaldi[28]平台下完成,端到端实验都在开源的 ESPnet(end-to-end speech processing toolkit)[29]端到端语音开发平台下完成，传统的模块化模型在 Kaldi[28]平台下完成。在所有的实验中，对于音频数据，都采用 1.6 KHz 的抽样频率，80 维的 FBANK 特征向量，每帧 25 ms，帧移 10 ms。

2.3　网络结构及参数

本文使用的网络，使用 4 层的多头注意力，每层输入注意力的维度为 256，前向全连接层的输入特征维度为 256，隐藏特征维度为 2048，双向长短时记忆单元数为 2048。联合训练参数 λ 为 0.3，随机丢失激活细胞率为 0.1，标签平滑为 0.1。集束搜索宽度为 10，语言模型占比重 0.1，epoch 次数为 50。

2.4　评价指标

本文在数据堂的开发集和测试集上评价实验结果，采用字错率(Character Error Rate，CER)和句错率(Sentence Error Rate，SER)作为评价指标，字错率即为了使识别出来的词序列和标准的词序列之间保持一致，需要进行替换、删除或者插入某些词，这些插入(I)、替换(S)或删除(D)的词的总个数，除以标准的词序列中词的总个数的百分比。句错率中，如果一个词识别错误，那么这个句子被认为识别错误。即:

$$\mathrm{CER}(\text{字误率}) = \frac{S+D+I}{N} \tag{17}$$

$$SER(\text{句错率}) = 1 - \frac{\text{正确句数}}{\text{总句数}} \tag{18}$$

2.5　对比实验

本文实现了 GMM-HMM 模型，时延神经网络模型(TDNN)和 kaldi 的链式模型三个传统模型作为对比实验，这三个传统模型也是数据堂开源的在带口音普通话语音数据集中表现最好的三种模型，采用 kaldi[28]工具包中 aidatatang 的方案(https://www.datatang.com)。

高斯—隐马尔可夫模型(GMM-HMM)：经过特征提取之后的语音特征序列，在忽略时序的条件下，研究者发现高斯混合分布非常适合拟合这样的特征序列，因此 GMM 被整合到 HMM 中，拟合基于状态的输出分布，提出了 GMM-HMM 模型。

时延神经网络模型(TDNN)：时延神经网络中可以对长序列的具有依赖时序的序列建模，采用二次采样的方法减少计算量，在训练上比同样对时序建模的循环神经网络要快，同时网络利用 i-Vector 特征，对自适应说话人和环境都有区分性，在带口音的训练中也有良好的表现。

链式模型(Chain Modle)：链式模型是 DNN-HMM 模型的一种，模型借鉴了 CTC 的思想，引入了空字符来吸收不确定的边界，使用正确序列的对数作为目标函数，在训练过程中要计算分母和分子的概率，并使用两者之间差异的倒数回传到网络中，链式模型同时可以有效地提高解码的速度。

基于多头注意力的模型[20,21]:多头注意力是基于自注意力机制的，自注意力模型是发生在序列内部，即在编码器内部学习语音特征的序列，在解码器内部学习到标签的特征，相较一般的注意力机制在序列问题上有更好的性能。

LSTM-RNN-CTC 模型[30]:结合了 LSTM 良好的序列建模能力，通过对不同地域的普通话口音做区分性训练，极大的提升了模型识别口音的能力。

3　实验结果及分析

3.1　传统模型与端到端模型对比

表 1 展示了传统的模块化自动语音识别系统和常见的端到端模型在多口音普通话语音识别上的效果，其中 MTL(Multitask Learning)参数代表联合训练中基于注意力机制的得分所占的比重，表 1 中混合端到端架构中的多头注意力机制模型采用 12 层的编码器-06 层的解码器架构。相较基于多头注意力机制的模型，混合模型有 10.0%-10.9%的字错率和 6.3%-7.1%的句错率的提升。在单独基于多头注意力的数据集中，大部分的识别错误集中在重口音普通话的说话人语句中。说明单独基于注意力机制的模型在拟合多个地区口音的识别上还存在严重的缺陷，尤其对于同一个字符的发音变异问题。相较传统的模块化识别模型，混合端到端模型展现出来了良好的性能，通常在识别率上领先基于时延神经网络模型(TDNN)和传统的 HMM-GMM 模型，略低于最先进的链式模型。但端到端的模型由于其良好的整体性可以很方便的达到对模型整体进行调优的目的。

表 1　带口音普通话语音在各不同模型中的识别率对比

Tab. 1　Comparison of the recognition rate of Mandarin speech with accents in different models

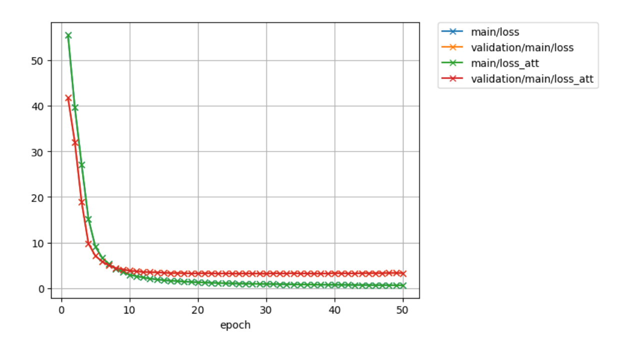
models	test	
	CER(%)	SER(%)
GMM-HMM	12.2	43.1
TDNN	7.1	31.2
CHAIN	5.6	26.1
Multi-Head	17.3	36.0
LSTM-RNN-CTC	15.1	32.7
CTC/Attention (Location+MOL(0.5))	10.5	39.0
CTC/Attention (Content+MOL(0.5))	10.9	39.7
CTC/Attention (Multi-Head+MOL(0.1))	7.3	29.7
CTC/Attention (Multi-Head+MOL(0.3))	6.4	28.9
CTC/Attention (Multi-Head+MOL(0.5))	6.5	29.7

如表 1，值得注意的是，随着多任务学习权重(MTL)的提升(0.0，0.1，0.3),识别率也有进一步的提升。如图 4 所示，在 CTC 损失权重在多任务训练中所占比重过大，模型的训练效果并不好，出现了严重的过拟合现象。这可能是由于 CTC 本身基于带条件的独立假设，在多口音的数据集中无法有效利用口音相似的特性，导致模型的收敛效果并不好。但在混合模型上，基于链接时序主义的方式仍然可以有效的帮助提升识别率，同时可以提升模型的收敛速度。在多任务学习的框架下，合理的配置设置训练权重，基于链接时序主义的方法可以强制性的进行特征和标签之间的单调对齐，对于基于多头注意力的模型的训练带来了很大的提升，同时也加快了基于多头注意力模型的收敛速度。

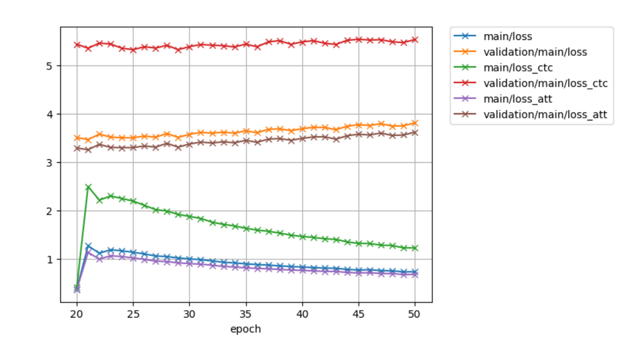
3.2　混合模型深度与随机失活率研究

为了进一步提升多口音普通话的识别率，加深编码器和解码器的网络层，并在浅层网络中逐步提升随机失活率分别为 0.0、0.1、0.5。

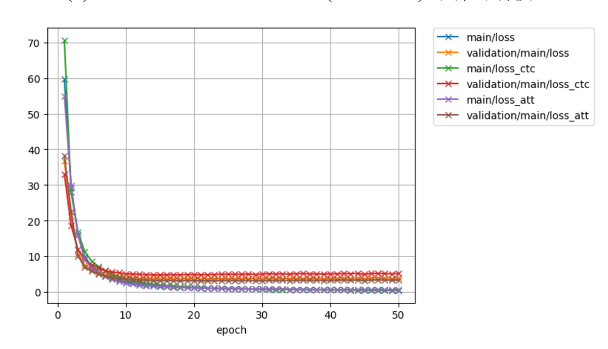
如表 2 所示，实验结果表明了本文提出的基于深层多头注意力机制和链接时序主义的模型，相较于传统的 GMM-HMM 模型，TDNN 模型，在识别率上都有较大的提升。相较常用的基于 LSTM-RNN-CTC 方法和基于多头注意力机制的模型，不仅在识别率有较大提升，在训练速度上也有一定的提升。同时随着层次的加深，字错率可以达到 5.6%，不需要字典就可以媲美目前表现最好的链式模型的识别率，同时模型简单、易优化。



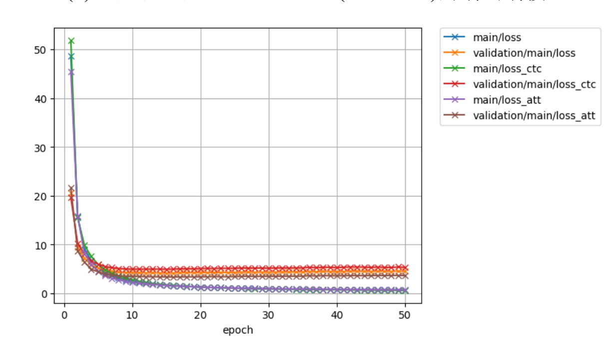
(a) 基于混合 CTC/Attention(MTL=0.0)的端到端模型



(b)　基于混合 CTC/Attention(MTL=0.1)的端到端模型



(c)　基于混合 CTC/Attention(MTL=0.3)的端到端模型



(d)　基于混合 CTC/Attention(MTL=0.5)的端到端模型

图 4　端到端模型训练损失图

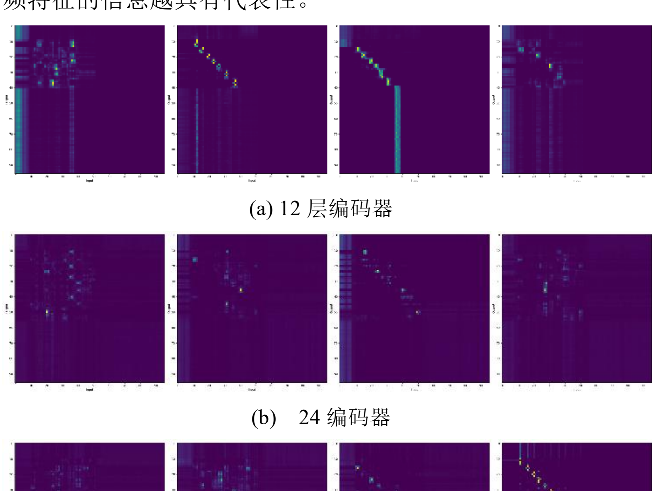
Fig. 4　End-to-end model training loss graph

表 2　深度与随机失活率研究

Tab. 2　Research on Depth and Random Inactivation Rate

models	test	
	CER	SER
Dropout Rate 0.0		
12Enc-06Dec	6.4	28.9
24Enc-24Dec	5.9	27.2
48Enc-48Dec	5.6	26.2
Dropout Rate 0.1		
12Enc-06Dec	6.3	28.5
Dropout Rate 0.5		
12Enc-06Dec	7.3	31.8

如图 5 所示，随着编码器层数的加深，编码器学习到音频特征的信息越具有代表性。



(a) 12 层编码器

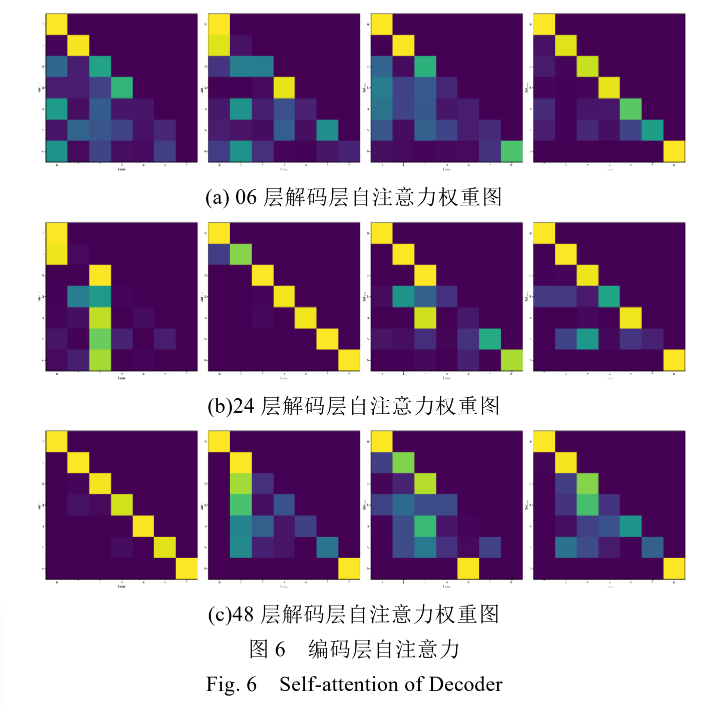
(b)　24 编码器

(c)　48 编码器

图 5　编码器自注意力

Fig. 5　Self-attention of Encoder

如图 6 所示，随着解码器层数的加深，加深了标签之间的相关度，在判定句结尾的静音片段上有显著的提升。相比浅层的网络，更少出现了识别少字的情况。



(a) 06 层解码层自注意力权重图

(b)24 层解码层自注意力权重图

(c)48 层解码层自注意力权重图

图 6　编码层自注意力

Fig. 6　Self-attention of Decoder

4　结束语

在本文中，提出使用链接时序主义(CTC)和多头注意力的编码器-解码器混合架构模型用于带口音的普通话语音识别。相较于传统的模块化模型需要大量的人工准备的资源，不用针对每一种方言准备语料和做模型自适应等复杂过程，只需要训练一个单一模型，不仅可以超过一般的传统方法的识别率，还可以方便的对模型的整体性能进行调优，可以达到最先进的模型的识别率。在提出的深度编码器-解码器层模型下，研究发现编码器的深度的增加可以更好的学到音频特征的信息，可以显著的提高带口音普通话的识别率，而解码

器的深度提升对于整个语音系统的提升较小，同时通过在浅层的网络中增加失活率，对于识别率也有微小的改进，后续的研究会通过改进失活方式，在深层的网络中进一步提升识别率。同时针对非常深层的网络训练速度较慢的问题，进行改进和提升。

参考文献：

[1] Huang C, Chen T, Chang E. Accent issues in large vocabulary continuous speech recognition [J]. International Journal of Speech Technology, 2004, 7 (2-3): 141-153.

[2] ZHANG D. xin (National Leading Group Office of the Teaching of chinese to Foreigners, Beijing 100083) [J]. My Point of View on the Standard of Newborn Words, 2000, 5.

[3] Ding G H. Phonetic confusion analysis and robust phone set generation for Shanghai-accented Mandarin speech recognition [C]// Ninth Annual Conference of the International Speech Communication Association. 2008.

[4] Elfeky M, Bastani M, Velez X, *et al.* Towards acoustic model unification across dialects [C]// 2016 IEEE Spoken Language Technology Workshop (SLT) . IEEE, 2016: 624-628.

[5] Wang Z, Schultz T, Waibel A. Comparison of acoustic model adaptation techniques on non-native speech [C]// 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP'03) . IEEE, 2003, 1: I-I.

[6] Zheng Y, Sproat R, Gu L, *et al.* Accent detection and speech recognition for shanghai-accented mandarin [C]// Ninth European Conference on Speech Communication and Technology. 2005.

[7] Huang Y, Yu D, Liu C, *et al.* Multi-accent deep neural network acoustic model with accent-specific top layer using the KLD-regularized model adaptation [C]// Fifteenth Annual Conference of the International Speech Communication Association. 2014.

[8] DeMarco A, Cox S J. Native accent classification via i-vectors and speaker compensation fusion [C]// Interspeech. 2013: 1472-1476.

[9] Chen M, Yang Z, Liang J, *et al.* Improving deep neural networks based multi-accent mandarin speech recognition using i-vectors and accent-specific top layer [C]// Sixteenth Annual Conference of the International Speech Communication Association. 2015.

[10] 张盼. 基于卷积神经网络的不同口音对话自适应识别研究 [D]. 重庆大学, 2018.

[11] Sak H, Senior A, Beaufays F. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition [J]. arXiv preprint arXiv: 1402. 1128, 2014.

[12] Yi J, Ni H, Wen Z, *et al.* Improving blstm rnn based mandarin speech recognition using accent dependent bottleneck features [C]// 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA) . IEEE, 2016: 1-5.

[13] Graves A, Jaitly N. Towards end-to-end speech recognition with recurrent neural networks [C]// International conference on machine learning. 2014: 1764-1772.

[14] Miao Y, Gowayyed M, Metze F. EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding [C]// 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU) . IEEE, 2015: 167-174.

[15] Chorowski J, Bahdanau D, Cho K, *et al.* End-to-end continuous speech recognition using attention-based recurrent nn: First results [J]. arXiv preprint arXiv: 1412. 1602, 2014.

[16] Bahdanau D, Chorowski J, Serdyuk D, *et al.* End-to-end attention-based large vocabulary speech recognition [C]// 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP) . IEEE, 2016: 4945-4949.

[17] Lu L, Zhang X, Renais S. On training the recurrent neural network encoder-decoder for large vocabulary end-to-end speech recognition [C]// 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) . IEEE, 2016: 5060-5064.

[18] Graves A, Fernández S, Gomez F, *et al.* Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks [C]// Proceedings of the 23rd international conference on Machine learning. ACM, 2006: 369-376.

[19] Cho K, Van Merriënboer B, Gulcehre C, *et al.* Learning phrase representations using RNN encoder-decoder for statistical machine translation [J]. arXiv preprint arXiv: 1406. 1078, 2014.

[20] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks [C]// Advances in neural information processing systems. 2014: 3104-3112.

[21] Kim S, Hori T, Watanabe S. Joint CTC-attention based end-to-end speech recognition using multi-task learning [C]// 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP) . IEEE, 2017: 4835-4839.

[22] Watanabe S, Hori T, Kim S, *et al.* Hybrid CTC/attention architecture for end-to-end speech recognition [J]. IEEE Journal of Selected Topics in Signal Processing, 2017, 11 (8): 1240-1253.

[23] Chorowski J K, Bahdanau D, Serdyuk D, *et al.* Attention-based models for speech recognition [C]// Advances in neural information processing systems. 2015: 577-585.

[24] Chorowski J, Jaitly N. Towards better decoding and language model integration in sequence to sequence models [J]. arXiv preprint arXiv: 1612. 02695, 2016.

[25] 胡章芳, 徐轩, 付亚芹等. 基于 ResNet-BLSTM 的端到端语音识别 [J/OL]. 计算机工程与应用: 1-8 [2020-03-26]. (Hu Zhangfang, Xyu Xuan, Fu Yanqin, *et al.* End-to-end speech recognition based on ResNet-BLSTM [J/OL]. Computer Engineering and Applications: 1-8 [2020-03-26]) http://kns. cnki. net/kcms/detail/11. 2127. TP. 20191014. 170 6. 010. html.

[26] Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need [C]// Advances in neural information processing systems. 2017: 5998-6008.

[27] Pham N Q, Nguyen T S, Niehues J, *et al.* Very Deep Self-Attention Networks for End-to-End Speech Recognition [J]. arXiv preprint arXiv: 1904. 13377, 2019.

[28] Watanabe S, Hori T, Karita S, *et al.* Espnet: End-to-end speech processing toolkit [J]. arXiv preprint arXiv: 1804. 00015, 2018.

[29] Povey D, Ghoshal A, Boulianne G, *et al.* The Kaldi speech recognition toolkit [C]// IEEE 2011 workshop on automatic speech recognition and understanding. IEEE Signal Processing Society, 2011 (CONF) .

[30] Yi J, Wen Z, Tao J, *et al.* CTC Regularized Model Adaptation for Improving LSTM RNN Based Multi-Accent Mandarin Speech Recognition [J]. Journal of Signal Processing Systems, 2018, 90 (7): 985-997.